said matching step includes exactly matching a single input phoneme of a subset at the top level of the hierarchical database but only best approximating a match at the lower level(s) of the hierarchical database.--

REMARKS

Reconsideration of this application is respectfully requested.

The Examiner is respectfully requested to reconsider the "finality" of the outstanding Official Action.

The present continuation-in-part application was filed with ten original claims including claims earlier presented in the first-filed parent application. However, applicant's intention was to issue only a single patent and thus the parent application was permitted to go abandoned. However, the present CIP application was subjected to a restriction requirement on March 31, 1997, and the applicant happened to elect to proceed with the patentably distinct claims 1 - 4 that had previously been pending in the parent application. Divisional applications (08/796,818 and 08/844,859) have since been filed to pursue the other non-elected groups of patentably distinct claims.

Clearly, if applicant had been permitted to pursue all of the additional new claims presented in this single application, then it would have been improper to make the first substantive Official Action a "final" Official Action. Under the circumstances where applicant has instead been forced, after abandonment of the parent application, to file two separate divisional applications to cover the additional claims (thus incurring <u>four</u> filing

fee costs), and thus, only under this changed circumstance is left with claims that were also being pursued in the parent application, it seems unfair and contrary to the intended spirit of the relevant regulations and MPEP processes to make the first Official Action a "final" Official Action.

Therefore, in the hope that the Examiner will agree to withdraw the "finality" of the first substantive Official Action in this case, in addition to formal amendments made to the specification and original claims 1 - 4, additional claims 11 - 14 have also been added. Entry of these additional claims into the present application (even under the provisions of 37 C.F.R. §1.116) is respectfully requested in view of the above-noted circumstances.

The specification has been amended above so as to bring it within the more typical guidelines mentioned by the Examiner.

By a separate (attached) letter to the Chief Draftsperson, drawing amendments are requested as shown in red so as to make it clear that the claimed invention (including the method and database) are schematically depicted in the drawings in accordance with 37 C.F.R. §1.83(a). Subject to the Examiner's approval and a notice of allowance, suitably corrected substitute formal drawings will be timely filed.

In response to the formality objections to claims 1 - 4, these claims have been extensively amended so as to insure proper antecedent basis, grammar, punctuation, and

the like. In particular, all of the Examiner's stated reasons for objection are now believed to have been obviated.

Accordingly, all outstanding formal grounds of objection/rejection are now believed to have been overcome.

The rejection of claims 1-4 under 35 U.S.C. §102(b) based on Jacks et al ('941) is respectfully traversed.

Claim 1 relates to a precise state in long and complicated process of synthesizing speech, namely, the conversion of phonemes into digital waveform signals. A great deal of preliminary processing is needed to produce the text in phonemes. There will also be subsequent processing of the waveform in both digital and analog formats before an acoustic output is produced. The claimed invention is not concerned with either the preliminary or the subsequent processing, but instead, relates to the critical step where phonemes are initially converted into digital waveform signals.

Although claim 1 has now been altered from its original European 2-part format, the first part of the claim admittedly does still read on Jacks et al. However, the second part of the claim (relating to the "extended digital waveform") is an important novel and patentably distinct feature. If the second part is compared with Jacks et al, the differences should be much clearer.

Claim 2 specifies that windows are compared. Claim 3 specifies that the windows are five phonemes long.

Claim 4 defines the input section of the database as being "hierarchical". The top level of the hierarchy is mentioned in item (i) which comprises selecting an exact match for the central (i.e., the third of the five phonemes in a given window) phoneme. Levels (i) and (iii) of the hierarchy are required to select a "best match" (i.e., where as in the usual case, an exact match does not exist).

The "extended digital waveform" may be provided by an individual speaking text to produce a digitized speech signal. The subsequent speech synthesis of other arbitrary input phoneme text is good enough that the "Talker" can be recognized. Depending on the age and/or sex of the "Talker", the machine-created audible speech will be recognizable as a man, or a woman, or a child. One case which was published in the British national press concerned a young girl who for medical reasons was not able to talk and she wanted to be able to type onto a keyboard and have the words spoken in a voice appropriate to a little girl of her own age. Andy Breen is named as the inventor on this case, and his daughter, therefore, recorded the "extended digital waveform" for this intended usage. When the resulting machine is used, it sounds like a little girl. The "machine" once participated in a BBC broadcast in which very good speech is synthesized.

There are many things which contribute to this good quality, but the extended digital waveform is a key feature. On her particular implementation, Andy Breen's

daughter can be recognized. The only thing which she contributed was the extended digital waveform and it, therefore, must be the waveform which makes her recognizable.

The extended waveform may include, for example, about two hundred sentences with about 40 phonemes per sentence, and the recording would last for about five minutes. (It would take much longer to record the text because the "Talker" makes mistakes and it is necessary to re-record. This is rather like a Hollywood film where it takes one day to make much less screen time.) At the top of page 5 of the specification it says that the exemplary extended text may be about 1000 - 1500 phonemes long and its lasts about two or three minutes.

In addition, there is an equivalent grapheme text which the "Talker" uses to make the recording. Linguists calibrate the phoneme text with a time parameter, e.g., milliseconds, so that selecting any one phoneme (or a string of phonemes) from the phoneme text enables the corresponding or "equivalent" waveform to be selected.

Jacks et al teaches that English can be represented with only 50 different phonemes. Thus, applicant's comparatively-extended waveform of more than 1,000 phonemes gives an average repetition of many times for each phoneme (Clearly, some phonemes will occur more frequently than others.).

Applicant's method comprises best-matching successive portions of an input text with portions of the "two hundred sentences". This selects portions of the extended waveform and the selected portions are built up into the final output. As is specified in

claim 4, the selected portion is either one, two or three phonemes long. This is clearly a very small fraction of a text which is thousands of phonemes long.

As noted above, individual phonemes occur many times in the extended waveforms. When one phoneme is matched, there are, on average, many matches. Because there is a choice, the method can take into account the preceding and succeeding phonemes. In the preferred embodiment (e.g., claim 3), the method takes into account the two phonemes before and the two phonemes after the matched central phoneme. In other words, the method is context related, and this gives better results. More sentences in the extended waveform (i.e., a longer or more extended digital waveform) would give more choices and even better results.

The extended database text is a key feature. As the source text gets longer, the number of phonemes in the available fixed "alphabet" necessarily remains constant and, therefore, the number of repetitions and contexts must increase. Good results depend on selecting a waveform sequence from among a variety of contexts. An extended waveform database inherently provides the variety which is needed.

Although single phonemes are most important, the occurrence of longer strings is also relevant. There are 2500 diphones so there is repetition even of diphones in a text of only 8000, for example. One might like to pick out triphones (i.e., strings of 3 phonemes) but, based on 50 phonemes, there are 125000 possible triphones. Clearly, these cannot be accommodated in a source text of only a few thousand phonemes. Claim 3 mentions

strings of 5 phonemes and there may be over 1 billion such strings. Clearly, these are not all accommodated in the exemplary embodiment. However, claim 4 specifies utilizing best available matches and one can therefore pick out strings of 5 which provide a "best match". More specifically, the claimed method picks EXACT matches for monophones and then selects diphones and triphones from best available matches which are 5 phonemes long. It is the use of an extended waveform which gives a wide choice and good results.

Applicant is not aware of any prior art which teaches the use of an extended waveform and selection of very small fragments, e.g., less than 0.1%, to make up the synthetic speech.

In the absence of an extended waveform, it is not possible to include the other claimed features of the invention. Even the concept of a "best match" is excluded from the prior art without an available extended waveform being available.

Context is important to obtain good synthetic speech. To the uninitiated a phoneme may be thought to represent a specific sound and, more specifically, to always represent exactly the same waveform. This is by no means correct. When human babies learn speech, we learn words and phrases and sentences. For convenience of pronunciation, we modify the sounds by what comes before and afterwards. Linguists find it convenient to use a single phoneme for a particular sound, but there are many modifications depending upon the context. It might be possible to synthesize speech

based upon an invariant waveform for each phoneme, but the results would not be very good because the context is neglected.

The sound of speech depends substantially upon the transitions between phonemes. Diphones provide a waveform for two adjacent phonemes and this includes the transition between them. Diphones can, therefore, give a better result than a system which only uses single phonemes. The Examiner may consider that a diphone takes context into account because it represents each of its individual members in the presence of the other, but this is not a context-related DIPHONE because the context of the DIPHONE is not taken into account. A diphone changes its waveform with its context and applicant's invention takes the context of the diphone into account. These comments relating to diphones apply to the transition contained in the diphone.

Jacks et al '941 describes "preliminary processing", i.e., grapheme to phoneme conversion. This is not part of the claimed invention, and this portion of Jacks is therefore irrelevant.

For example, column 1, line 65, to column 2, line 19, of Jacks et al. clearly is concerned with the generation of a phoneme text from a grapheme input. This quote mentions,

> identifying clauses within text sentences by locating punctuation and conjunctions;
>
> analyzing the structure of each clause by locating key words;

> converting the sentence structure thus detected in accordance with the standard rules of grammar, into prosody information, i.e., inflection, speech and pause data.

These may all be part of "preliminary processing" but that does not suggest any extended waveform library.

Column 2, line 20, of Jacks et al specifically states "when the proper phonemes sequence has been determined". This sentence by itself is conclusive that the description is concerned with generating the phoneme text and not with converting the phoneme text into digital speech waveform.

The same applies to column 4, line 28, to column 5, line 446. Column 5, line 42 states that the result consists of phoneme codes, i.e., nothing more than applicant's admittedly-old preliminary processing. The Examiner mistakes statements relating to Jacks grapheme to phoneme conversion with Jacks later <u>conventional</u> phoneme to waveform to conversion. This is a fundamental error.

Jacks gives a good description of his phoneme to waveform conversion. Column 1, lines 51 to 53, refers to "a limited number of very small digitally encoded waveforms". This passage by itself makes it clear that Jacks uses SHORT WAVEFORMS and that is clearly a fundamental difference from applicant's claim 1 which specifies an EXTENDED waveform.

At column 1, lines 57 to 60, Jacks refers to smooth transitions from one phoneme to another with a minimum of data transfer. This makes it clear that Jacks utilizes not

only single phonemes, but also transitions and, again, Jacks emphasizes "a minimum of data transfer" which is fundamentally different from using an EXTENDED waveform.

This fundamental difference is further emphasized by Jacks at column 2, lines 32 to 43, which mentions that the "speech segments are very small". There is a further reference to "because of the extreme shortness of the speech segments" at column 3, line 61. Jacks is now talking not only about "shortness" but "extreme shortness".

Jacks discusses transitions at column 3, lines 20 to 30. He states that transitions require a large amount of memory and substantial memory saving can be accomplished by calculation. He says that only two segments are required because the transition can be calculated from the two segments.

It is clear that from Jacks' description, especially since he emphasizes the small amount of memory which is needed, that Jacks only provides a single waveform for each phoneme and single waveform for each transition. In other words, Jacks does not take context into account at all.

Jacks gives good description of his method at column 5, lines 60 to 68. Specifically, he gives 9 steps to convert the phonemes representing the word "speech" into a waveform. He acquires a waveform for each transition, a waveform for each phoneme and he concatenates these to produce the desired output.

At column 6, lines 12 to 22, Jacks mentions that there are "50-odd phonemes" and "2,500-odd possible transitions" from one to the other. Jacks explicitly states that he

provides 50 waveforms for the phonemes (i.e., one for each) and 2,500 waveforms for the transitions (i.e., one of each). In other words, Jacks provides a total of about 2,550 waveforms. Some, but not all, of these waveforms are stored explicitly, but they can all be acquired (by calculation) from stored data. jacks is very concerned with minimizing the amount of storage. Thus, he calculates transitions from small amounts of data and he even uses backward running transitions in order to save storage space.

Thus, a fundamental difference between Jacks et al and applicant's claimed invention is that he does not use an extended waveform. Because he does not use an extended waveform he has no alternatives and he cannot make selections. He only utilizes single phonemes and transitions and, therefore, he cannot find best matches to five phoneme sequences.

The Examiner refers to column 4, lines 60 to 63. This is part of the grapheme to phoneme conversion, and it has nothing to do with the claimed invention. The Examiner includes the word "parsing", however, it is clear that "parsing" has nothing to do with phoneme to waveform conversion.

The Examiner mentions "retrieving a digital waveform linked to the input". Column 6, lines 12 to 21, 38 to 47, 61 to 66, and column 7, lines 1 to 4, are mentioned. Of course, any conversion system must have an input and an output, but the input and output described by Jacks bears no resemblance to the ones specified in claim 1. Jacks does not have an extended waveform; in fact, he emphasizes that he has very short

waveforms. Since he has not got an extended waveform, he cannot select portions of it. Jacks is fundamentally different from the claimed invention.

Joining digital segments together is a feature of all synthetic speech. The claimed invention is concerned with how the digital segments are obtained, and in this aspect the invention is nothing like Jacks. The Examiner refers to column 7, lines 1 to 3, and he alleges that these disclose "an extended waveform". This is clearly not correct because Jacks emphasizes the "extreme shortness" of his segments., and he has no "extended waveform". Furthermore, Jacks does not refer to "a location parameter". He refers to the address at which the relevant segment may be obtained. Jacks also emphasizes that the segments are "extremely short".

The Examiner refers to column 6, lines 61 to 64 and alleges that this discloses "beginning and ending location parameters". This is not true. The claimed invention uses beginning and ending parameters because it uses an extended waveform. In order to select a portion (whether it be long or short) from an extended waveform, it is necessary to know where the chosen portion starts and where the chosen portion ends. Therefore, the applicant uses beginning and ending location parameters. Jacks does not have an extended text and, therefore, he cannot select portions of it. He has short segments and he merely selects the appropriate short segment. When he selects the segment he uses all of it. Jacks has no beginning and ending parameters.

The Examiner mentions claim 9 step (g) which refers to storing sequences of digital data representing segment blocks corresponding to particular phonemes and the transitions therebetween. This apparently means particular phonemes and particular transitions. This means that each phoneme has a waveform (but only one) and each transition has a waveform (but only one). This is not "context related" because a particular phoneme always gets the same waveform whatever its context and, likewise, a particular transition always gets the same waveform whatever its context.

The Examiner uses the phraseology "thereby defining the database as a context based database". Thus, it appears that the Examiner regards Jacks as "context based" because he uses transitions as an element. The precise nature of a transition is affected by what comes before the first phoneme and what comes after the second phoneme. It is clear that Jacks does not use context related transitions.

The Examiner says that Jacks teaches "comparing windows of an input signal with windows of the input section of the database" as mentioned in our claim 2. This statement is incorrect. Jacks does not have an extended waveform and, therefore, he does not have a text corresponding to an extended waveform.

In his next paragraph the Examiner alleges that column 5, lines 50 to 56, discloses windows with a length of 5 phonemes. This is clearly wrong. Jacks has nothing in his database as long as 5 phonemes. He has equivalents for each particular phoneme, but he only has equivalents for SINGLE phonemes. These, clearly, are not strings of 5. He

provides a short waveform corresponding to the transitions between the two elements of a diphone, but this is substantially less than the diphone. In any case, it must be less than a string of 5 phonemes. Furthermore, claim 3 makes it clear that a string of 5 phonemes is selected from an even longer text. Clearly, Jacks has nothing that remotely resembles this.

The Examiner makes comments regarding applicant's claim 4 which are not understood. As specified in claim 4, the method breaks the input text into windows and matches these, by a search procedure specified in the claim, against windows of a text corresponding to the extended digital waveform. Each window has 5 phonemes. Jacks does not do this or anything like it.

Claim 4 specifies that at least one (i.e., phoneme number 3 in the window) is matched exactly against a phoneme in the extended text and the other phonemes are matches as closely as possible. Applicant's 200 sentences, for example, contain phonemes in many different contexts and, what is achieved in the method of claim 4, is the selection of the most appropriate window.

Having identified the best window, applicant's method selects 1, 2 or 3 phonemes therefrom and retrieves the corresponding portion of the extended waveform. These are used to create the output. This gives very natural speech which Jacks can not easily achieve.

As a final remark on Jacks et al, it is noted that he is particularly concerned with using the smallest possible amount of storage space. That is not part of the claimed invention. Modern computers have enough storage space to hold the extended waveform and the search technique specified in claim 4. Applicant does not wish to necessarily minimize storage space because applicant is more concerned about getting a natural result.

Attention is also directed to new claims 11 - 17. Claims 11 - 14 are similar in many respects to claims 1 - 4. Claims 15 - 17 provide yet another approach to claiming use of applicant's novel and non-obvious extended digital waveform database.

Accordingly, this entire application is now believed to be in allowable form, and a formal notice to that effect is respectfully solicited.

Respectfully submitted,

NIXON & VANDERHYE P.C.

By: Larry S. Nixon

Larry S. Nixon
Reg. No. 25,640

LSN:maw
1100 North Glebe Road, 8th Floor
Arlington, VA 22201-4714
Telephone: (703) 816-4000
Facsimile: (703) 816-4100